DOCUMENT RESUME

ED 434 403 EA 030 065

TITLE Assessing Excellence: A Guide for Studying the Middle Level School.
INSTITUTION National Association of Secondary School Principals, Reston, VA.
ISBN ISBN-0-88210-218-4
PUB DATE 1997-00-00
NOTE 62p.; For the first two books in the series, see ED 263 679 and ED 292 209.
PUB TYPE Guides - Non-Classroom (055)
EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Educational Administration; *Educational Assessment; Evaluation Methods; *Formative Evaluation; *Middle Schools; Models; Program Evaluation

ABSTRACT

This volume is the third in a projected six-volume series called "Agenda for Excellence at the Mid-Level." It describes a process that school administrators can use to conceptualize, plan, execute, and utilize the results of a school assessment, allowing them to take a snapshot of their school. It outlines a set of alternative procedures to be used to assess the middle-level school so that judgments can be made about the way the school is operating and whether it is achieving desired outcomes. The text proposes a process not only for conducting the assessment, but for making judgments about the findings and turning those judgments into action plans. The process itself is designed to make use of existing, internal resources and does not require, or even recommend, the use of external systems or consultants. It outlines the purposes of evaluation and describes an evaluation model, providing an example of that model. It looks at prepared evaluation systems and provides insights on data analysis and interpretation. It shows how to make use of questions and analytical frameworks and how to deal with unanticipated findings. It also discusses recommendations to make following the evaluation, notes on disseminating the results, and pitfalls to avoid. (RJM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
R. Mahaffey
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

ED 434 403

ASSESSING EXCELLENCE: A GUIDE for STUDYING THE MIDDLE LEVEL SCHOOL

NASSP's Council on Middle Level Education

EA 030065

BEST COPY AVAILABLE

ASSESSING EXCELLENCE: A GUIDE FOR STUDYING THE MIDDLE LEVEL SCHOOL

National Association of Secondary School Principals
1904 Association Drive • Reston, Virginia 22091

NASSP's Council on Middle Level Education

Alfred A. Arth, *Professor, University of Wyoming*

Sherrel Bergmann, *Professor, National College of Education*

J. Howard Johnston, *Professor, University of Cincinnati*

John H. Lounsbury, *Editor, The Middle School Journal of the National Middle School Association*

Conrad F. Toepfer, Jr., *Professor, State University of New York, Buffalo*

George E. Melton, *Deputy Executive Director, National Association of Secondary School Principals*

Copyright 1988
Reprinted 1997
National Association of Secondary School Principals
1904 Association Drive
Reston, Virginia 22091
(703) 860-0200
ISBN : 0-88210-218-4

Contents

Foreword iv
Preface vi

Study Your Middle Level School 1
An Overview 3
Purposes of Evaluation 4
An Evaluation Model 15
An Example 27
Prepared Evaluation Systems 33
Data Analysis and Interpretation 35
Use of Questions and Analytical Frameworks 42
Dealing with Unanticipated Findings 43
Making Recommendations 44
Dissemination of Results 49
Some Pitfalls 51
A Final Note 53
References 53

Foreword

In the past decade, an increasing number of schools have changed their grade configuration-i.e., from 7-9 to 6-8-or changed their name from junior high school to middle school. However, not all schools that have changed their grade configuration or name have, indeed, become effective middle level schools. In fact, some school: that have grade configurations other than 6-8, or that call themselves junior high schools, have more effective programs than schools that call themselves middle schools. Effectiveness is more than what's in the name or in the grade configuration.

The success of a middle level school lies, rather, in the implementation of a program that successfully meets the needs and characteristics of young adolescents. Such implementation results in a school that demonstrates significantly positive student achievement and school climate.

Recent research indicates that there are essential elements in successful middle schools, and that effective middle schools do not follow a "checklist," but rather consistently work hard to implement programs that they learn improve student achievement and school climate. Schools are becoming increasingly accountable to their local communities and states.

Middle schools, especially, have come under attack, often because of their inability to demonstrate how the programs they are implementing are contributing to student achievement and development. Assessing Excellence: A Guide for Studying the Middle Level School, originally published in l988 as part of the series Agenda for Excellence, with J. Howard Johnston as the primary author, is still an appropriate guide to examining your own program so you cannot only continue to improve, but also share with your public exactly how you

are contributing to your students' education. We hope it will be a valuable tool for you and your staff to assess your own school's excellence.

Timothy J. Dyer
Executive Director
NASSP

Susan Galletti
Associate Executive Director,
Director of Middle Level Services
NASSP

Preface

Assessing Excellence: A Guide for Studying the Middle Level School is the third in the projected six-volume *Agenda for Excellence at the Middle Level* series. In it, the NASSP Council on Middle Level Education, with J. Howard Johnston serving as project director and major author, proposes a plan for assessing a middle level school or a part of it. The plan can serve as a general model for evaluation and an introduction to thinking about evaluation as a social-political or a scientific enterprise.

This volume is not an evaluation textbook, a measurement book, a statistics refresher course, or an evaluation manual. It is a description of a process that may be used to make a significant yet relatively easy assessment of a school. Building upon *An Agenda for Excellence at the Middle Level* and *Developing a Mission Statement for the Middle Level School*, this volume continues the efforts of NASSP to assist its members in the improvement of their middle level schools. On behalf of those members, I acknowledge the debt we owe the NASSP Council on Middle Level Education for the major role they have played throughout this process.

George E. Melton
Deputy Executive Director
NASSP

Study Your Middle Level School

There is no process more complicated yet so often used as evaluation. We make judgments of quality virtually every moment of every day. We judge the quality of a meal, the quality of service at a local gas station, the entertainment value of a television program, or the relative worth of a purchase. In fact, most of the decisions we make are based on evaluations we have conducted.

Sometimes those evaluations are very formal and systematic, such as comparing two cars before buying one, or determining whether or not we should purchase a new home. Most evaluations, though, are spontaneous, impressionistic, and very informal, such as the evaluations we make about whether a new movie is worth seeing, or if we should try a new restaurant.

In every case, however, an evaluation serves one purpose: it helps us to determine the *value* of something. More important, that value is always expressed in terms of some *cost*. The question it seeks to answer is whether or not the activity, the product, or the event is worth the personal resources it will consume.

The evaluations we perform in schools are no different. We are trying to assess the value of student work, the effect of a new reading program, the impact of our counseling program on school dropout rates. In every instance, we must ultimately judge the worth of the evaluation object.

Thus, evaluation can be thought of as having two dimensions: assessment and judgment. For the purposes of this monograph, assessment is a process used to determine the status of some event, product, or program. In other words, assessment answers the question, "What's going on here?"

Judgment follows assessment. It is the process by which we

assign a value to the event, product, or program. It answers the question, "What is it worth to us?"

Assessment processes can be standardized. We can be certain that we gather data in the same way in different settings or at different times. We can use virtually identical procedures and instruments. We can carefully control our analytical procedures so that they are comparable from group to group and time to time. Just as with a recipe for a cake, we can be fairly certain that using the same process each time will produce similar results.

Judgment, though, is entirely local. Only a school or a community can judge, appropriately, whether the results they get from their assessment are what they expect, want, or need. To continue our analogy, the process prescribed in the recipe may produce a chocolate cake, but only the consumer of the cake can judge how satisfying it is for his or her appetite.

This monograph is devoted to *assessment*. It outlines a set of alternative procedures to be used to assess the middle level school so that judgments can be made about the way the school is operating and whether it is achieving desired outcomes. The process that is outlined here ***will not make decisions for the evaluators***. It will not ascribe value to the results of the assessment.

Instead, we will propose a process not only for conducting the assessment, but for making judgments about the findings and turning those judgments into action plans.

This is not designed to be a text on school evaluation. There are many good books on the subject. This monograph is designed to help school administrators conceptualize, plan, execute, and utilize the results of a school assessment. It describes a way to take a snapshot of your school.

Perhaps most important, this process is designed to make use of existing, internal resources. It does not require (or even recommend) the use of external systems or consultants. It does recommend that the assessment be conducted as a broad-based, highly participative activity that includes as many peo-

ple from the school and community as possible.

It also suggests that evaluations need not be massive, school-wide tasks that consume enormous amounts of time and energy. In fact, often it is the very narrowly focused assessment which answers only one question ("How much participation are we getting in our intramural program?") that yields the most useful information.

An Overview

An evaluation is designed to answer a series of questions. This monograph addresses each of these questions in turn:

- What are we evaluating? How precisely can we specify the object of our evaluation?
- What is the purpose of our evaluation: diagnostic, program monitoring and accountability, program impact, or cost-benefit?
- Who are the stakeholders in this evaluation? Who is going to be affected by the fact that it was done? How can I involve them in the evaluation?
- What are the goals of the program?
- What will indicate that the goals are being met?
- What kind of data are available on each of our indicators?
- How are the data to be analyzed?
- How can we get consensus on results, conclusions, and recommendations?
- What recommendations are supported by our data?
- What is the best way to disseminate our results and allow feedback?

Purposes of Evaluation

Most often, evaluations are conducted to meet a relatively specific need. Often, evaluators are trying to determine if some kind of intervention is actually needed. (Do we need an after-school activity/intramural program?)

In other cases, an administrator may want to know if a program is operating as it was supposed to. (Are students participating in the program?)

Just as often, we are interested in knowing the effect a program is having on people. (Are more students continuing to participate in athletic activities beyond middle school?)

And, finally, we are usually concerned with the efficiency of a program, or its cost relative to its benefits. (What is the after-school program costing us per student participant?)

Clearly, each purpose is driven by a different set of questions, and, as might be expected, slightly different procedures.

Most evaluations fit one of four categories: diagnostic evaluations, monitoring or accountability evaluations, impact evaluations, and efficiency or cost-benefit evaluations. Each type is driven by different types of questions and assumptions.

Diagnostic evaluations are designed to help us understand the discrepancies that exist between some ideal state and where we actually are. In other words, diagnostic evaluations help us see where an intervention might be needed in order to bring about a more desirable set of circumstances for our clients.

Diagnostic evaluations tend to focus on the needs of our clients, most often students and their families. Generally, a diagnostic evaluation is undertaken in response to a need identified by some stakeholder. For example, a group of minority parents may feel that not enough successful minority role models are being provided for their children. Their ad-

vocacy of this position may lead the school to undertake a study of its staffing practices, reward and incentive programs, grouping practices, curriculum materials, library holdings, and the like.

Usually, a diagnostic evaluation is undertaken because enough informal evidence exists to suggest a need may actually exist, and it will benefit the school and its clients to have a clear picture of this need in order to determine if an intervention is necessary or reasonable.

The questions associated with diagnostic evaluations tend to be as follows:

1. *Who are the targets?*

   These may be individuals or, just as likely, groups of people who are served by the institution. Usually, these are described in demographic or social terms (race, family status, age grouping, membership in a group), physical terms (where they are housed, the school they attend), or geographical or political terms (residence, region of country, citizenship, etc.).

2. *How many need service?*

   Once the target group has been identified, it is important to know how large the group is. Not only is it important to know how many people are likely to be served in absolute terms (is it 27 or 2,700?), but what percentage of the total institutional population is in need?

3. *Is it a need or a demand?*

   This is a judgment call. A new state mandate to teach foreign language to everyone who wishes to enter a state university places a legal obligation on the school to offer foreign language. On the other hand, the presence of a foreign-owned manufacturing facility in a community may point to a need for more foreign language education in the community.

   Generally, the distinction between a need and a demand

can be made by asking, "What will happen if we do nothing?" If the answer is that the clients will suffer dire consequences (physical harm, legal difficulty, career termination, etc.) it is a demand; if the consequences are merely undesirable (fewer students enter top colleges, students score lower on achievement tests) it is probably a need.

With scarce resources, schools are more likely to respond to demands than to needs, unless some political measure is taken to elevate a need to a demand, as might be the case where a parent group lobbies successfully with the board of education to require that foreign language be taught to all students in all grades.

Monitoring and accountability evaluations stem from our need to determine whether a program is performing as promised. Usually, a monitoring evaluation is specified in the original design of a program as a guarantee that the program managers will run the intervention as approved by the decisionmakers. In fact, funding agents, whether they are external agencies or boards of education, usually insist that a monitoring evaluation be conducted as a condition of funding.

The questions that drive a monitoring evaluation are somewhat different from diagnostic questions. Monitoring evaluations assume that a need was identified and that some program was implemented to address that need. Therefore, the questions that drive a monitoring evaluation are very simple.

1. *Is the program serving the people it was designed to serve?*

   Are the members of the target population actually receiving the service, or is it being misdirected? Nothing infuriates a funding agency more than having resources earmarked for a specific population used to serve another group.

   If, for example, a school board approved the expenditure of funds for an after-school care program only to find that it was serving only affluent students who had

after-school transportation available to them and not adequately serving poorer students who had to rely on public transportation, the future of the program would be in jeopardy.

2. *Is the program being conducted as promised?*

   This question is at the heart of accountability. Are we doing what we promised we would do? Does the program contain the elements it was supposed to contain?

   An advisory program that turns out to be a genuine guidance experience for some students, a study hall period for others, and a free conversation time for still others is likely to be viewed with great skepticism by the decisionmakers who approved the implementation of the program. On the other hand, an intramural program that is broad-based enough to encourage handicapped students to participate is a good example of an intervention that is true to its original intent of increasing participation opportunities for all students in the school.

Impact evaluations help us determine if a program is achieving its intended outcomes. Since most programs are goal-directed, impact evaluations are designed to determine if our goals are being met, to what extent they are being met, and if the outcomes can, in fact, be attributed to the program.

Typically, impact evaluations are the most difficult to conduct, while, at the same time, they are usually the most important from the point of view of the client and the funding agent. In essence, impact evaluations answer the all-important question, "So what?"

Your intramural program may have a large number of participants, but so what? What is the ultimate benefit to these participants? What personal, educational, or social good is achieved by their participation? In other words, what is the effect of their participation?

Obviously, answers to these types of questions are elusive because the effects of any program intervention may not be

readily apparent. At the same time, it is not reasonable to expect program sponsors to pay for programs that promise only ephemeral results some time in the distant future. For these reasons, impact assessments must examine immediate effects, potential long-term effects, and unanticipated outcomes.

Assessing immediate and long-term effects is quite straightforward. Typically, these assessments are related directly to the goals of the program.

It is much more difficult to assess unanticipated effects, those effects that were not anticipated by the original program planners. The reason for the difficulty is clear: What makes the evaluator more likely to detect the unanticipated effect of a program than the original developer of a program?

Finally, it is important to determine if the outcomes that are observed can be linked, directly, to the program. Is it the program that produced the outcome, or some other set of circumstances and events that would have occurred anyway?

The questions driving an impact assessment, then, become considerably more complex than those that guide other evaluations.

1. *What were the goals of the program?*

   The specification of intended short and long-range outcomes is absolutely essential in conducting an impact evaluation. Often, these goals are clearly stated in some kind of planning document. Just as often, programs evolve without very clear goals.

   In the former case, it is a relatively simple matter to list the goals and identify, from them, outcomes that would be expected in light of those goals. If, for example, the goal of the intramural program is to increase participation in athletics in high school, a simple comparison of the high school participation rates of students who were in the program and those who were not will help to indicate achievement of that goal.

In cases where no goal statements exist, it may be necessary to begin the impact assessment with a study of the perceived goals of the program as articulated by the people who run it, participate in it, and pay for it. Such a process can be undertaken in a pre-evaluation study described in a later section of this monograph.

2. *What will indicate that the goals are being achieved?*

Generally, it is impossible to study goal attainment directly. If the goal of a mathematics enrichment program is to "improve student attitudes toward mathematics," it is clear that attitude improvement cannot be observed directly; some kind of behavior must indicate that improvement.

Thus, the evaluator may specify that attitude improvement, for the purpose of the evaluation, will be evidenced by changes in scores on a subject preference inventory and by observing that more students sign up for advanced mathematics in high school. We must infer that these behaviors are indicative of more favorable attitudes toward mathematics.

It is important that everyone who has a stake in the program be engaged in identifying and agreeing upon indicators of program goals. Otherwise, the evaluator may amass a body of data which no one accepts as indicative of the goals they originally planned.

3. *In what domains should I expect to see effects?*

Often, programs produce effects that were unanticipated. A student activity program may promote interest in rocketry and, ultimately, increase enrollments in astronomy courses. A writing program may stimulate the production of a school literary magazine; a community service program may have a strong public relations benefit for the school.

In each case, some of the impact of the program is outside predicted areas; thus, they might be missed in an

evaluation. If the evaluator specifies the domains in which a program may have effect, they are more likely to be detected in a careful program evaluation. This increases the likelihood that all relevant program effects will be uncovered and considered, especially the unanticipated spin-off effects that are often overlooked.

Cost-benefit evaluations help to determine if we are getting our money's worth out of a program or activity. They tell us how much it costs to achieve some unit of benefit. In the case of an intramural program, a cost-benefit study may be used to identify what it costs to run the program per participant. In this case, the cost of the program, in terms of the benefit we have adopted, can be decreased by increasing participation without increasing expenditures. Thus, serving more students with the same resources would enhance our cost-benefit position.

Formal cost-benefit evaluations are rare in education because our outcomes are often not easily measured. At the same time, informal cost-benefit judgments are made all the time in schools and in the communities they serve.

A school may decide that its transportation program is just too costly to maintain at current levels in a time of budgetary restraint. A community may vote "no" on a bond issue to build a swimming pool at the middle school because it costs too much in absolute dollar terms. Seldom are such decisions made in light of cost-benefit estimates.

Similar difficulties arise when we try to assess the effects of a new program, in reading for example, on the students experiencing it. If the program costs $100,000 per year, and the 100 students in it raise their scores by 1.3 grade level equivalents during the program, is the benefit great enough to justify the cost?

At the most basic level the benefit of the program can be expressed in the following way: 100 students raise their scores an average of 1.3 years, for a group gain of 130 years increase

(100 x 1.3). This gain cost $100,000, or $769 per year of increase in group scores. That sounds pretty expensive, but only because we have expressed our benefits in a single term.

Other indicators of benefit must be examined as well. First, what was the effect of success on the academic self-esteem of the learners, and what effect does that have on their subsequent performance in reading and in other subjects? Second, what effect does the increased reading competency have on the amount of reading the students do? Because they can now read better, will they read more? And if they read more, will their ability continue to improve without additional special intervention?

Finally, will this simple gain open up new opportunities to them? Will they score better on employment tests that lead to higher levels of technical training? These questions can hardly be answered in a time-constrained evaluation plan, but they may be benefits nevertheless.

In order to conduct an adequate cost-benefit study, a number of questions must be answered with some precision.

1. *What, exactly, are we trying to study?*

   In order to identify potential benefits, it is important to define as precisely as possible the exact program element being studied. Rather than examining the counseling program, for example, the evaluator should specify some particular facet of the program: academic counseling of high-risk students.

2. *What are the likely benefits of this program element?*

   The answer to this question should, again, be as explicit as possible. The list should also be exhaustive, including both the outcomes anticipated by the program's objectives, and those that become evident as the program operates. This step is absolutely essential; without it, very desirable potential benefits may be lost from consideration.

   As an example, in one dropout prevention program, the evaluator noted that the students' achievement scores im-

proved only slightly, but that their attendance improved dramatically. Since that was not one of the program's original objectives, its benefit would have been overlooked if the evaluator were not attentive to unanticipated outcomes.

3. *What does the program cost?*

This is one of the most difficult things to determine, since school programs are seldom discrete. In industry, the cost of manufacturing a car can be assessed quite accurately, since each piece consumes certain resources and can have a precise cost affixed to it. If the car were not built, there would be no need for the part and, thus, no cost.

In schools, we seldom have such specificity. Because people in schools tend to fulfill multiple functions, eliminating the function does not eliminate the cost associated with it. For example, it is unlikely that a new counselor would be added for the purpose of running an at-risk program. The responsibility for that particular intervention would simply be added to an existing counselor's duties, or would replace something the counselor would otherwise be doing. Eliminating the program would not eliminate the cost of the counselor's salary; it would simply allow him or her to be assigned elsewhere.

Most programs have three major elements: personnel costs, materials and equipment costs, and overhead costs. Personnel costs are salaries, wages, and benefits associated with the operation of the program; materials and equipment costs describe expenditures made for specialized equipment that is needed to run the program; overhead costs are the additional operating expenses (custodial service, wear and tear on facilities, etc.) incurred as a result of the program.

Because school programs are not discrete, it is usually necessary to assign portions of existing costs to the operation of a given program. An example illustrates the point.

To calculate the cost of the at-risk counseling program, we might use the following estimates:

Personnel Costs	
One-fifth of counselor's salary and benefits. ($38,000 x .20)	$ 7,600
Full-time aide for tutoring	12,000
Materials/Supplies	
Computer to track student progress	2,300
Group counseling materials (texts)	600
Miscellaneous materials/supplies	600
Overhead Costs	
Maintenance of special classroom (66 classrooms in building, $112,000 annual maintenance cost: cleaning, heat, lights, insurance, repair, etc.)	1,696
TOTAL COST	$24,796

This program costs $24,796 to operate for one year. If it serves 30 students, the annual cost per student is $826.53 per year.

4. *What is the relationship between costs and benefits?*

In many ways, this is the trickiest part of the evaluation. It implies that some value can be assigned to the benefit and that some criteria exist that allow us to determine if something costs too much or not.

In some cases it is enough to examine only the cost of

coverage. In this case, we know that it costs just over $826 per pupil to cover the at-risk population.

At other times, it is necessary to understand the cost of the program in terms of its impact. In the foregoing case, we may discover that the dropout rate for students not in the program is 22 out of 30, and for students in the program for two years, it is 9 out of 30. Thus, if all other things are equal, the program reduced potential dropouts in the group by 13, the difference between those who had the program and those who did not.

In basic terms, it cost the school system $49,592 to run the program for two years, with the effect that 13 students were prevented from dropping out. In other words, it cost $3,814.76 to prevent each of these dropouts.

5. *Was it worth it?*

The last question is most critical. Did the benefit justify the expenditure? In the preceding example, the answer is clear. Given what a dropout costs in terms of lost tax revenue, additional welfare expenditures, and greater likelihood of crime and drug and alcohol abuse, $3,814.76 is a small price, indeed.

Unfortunately, most outcomes are not so clear-cut. All cost-benefit evaluations end with a judgment about whether or not the expenditure is worth the return. That is what makes the answer to Question 4 so critical. The benefits must be specified in terms that are clearly relevant to the consumer of the evaluation.

Some individuals may think that a cost of $826 per pupil per year is outrageous for a special program. Only when the benefit of the program is measured against the thousands of dollars that are wasted when we squander the life of a dropout is the true value of the program evident. Ultimately, it is the job of the evaluator to help the consumer ascertain the value of a program in terms that the consumer can understand.

An Evaluation Model

A model is a representation of reality that helps us to understand phenomena that are too large, too small, or too complex to comprehend all at once. Evaluations are usually complex, too complex for a single recipe to be useful in guiding them. In addition, the evaluation needs of different schools and different communities are unique and special. An evaluation that comes from outside the local context is not likely to produce much useful information. First and foremost, evaluation is and should be a local process.

This model identifies critical issues that must be addressed as an evaluation progresses. It is not a recipe; instead, it is a map through a successful evaluation plan. The issues to be addressed are stated as questions, the answers to which are the substance of the evaluation design and plan.

This model is based on several assumptions that must be stated clearly at the outset.

- Evaluation is a social process that depends upon agreement, compromise, ownership, and negotiation.
- Evaluations are credible only to the extent that all stakeholder groups have participated in them.
- Evaluations utilize events, people, and artifacts to determine if a change has occurred.
- Evaluations must conclude with value statements that judge the worth of something; thus, evaluations must acknowledge the value positions of participants in the process.
- Evaluations do not eliminate the need for judgment, they inform judgment.

First Steps

The most important part of any evaluation is to determine who is involved, and how. Two kinds of participants become particularly important: stakeholders and audiences.

Stakeholders are those individuals who have something to lose or gain as a result of the evaluation being conducted. Most often, these are the clients of a program—students and parents. However, stakeholders also include program providers, such as teachers, coaches, administrators, or contractors. A stakeholder is anyone who has a material interest in the effective operation (or discontinuation) of a program or program element.

Audiences are the groups for whom the evaluation results are intended. It is important to understand who the audience is so that the evaluation can address the kinds of concerns they are likely to have. A school board, for example, is more likely to be concerned about the costs and benefits of a program, while teachers working in the program may be more interested in the overall educational benefit of the operation. That doesn't mean that the focus of the evaluation should be determined by the narrow interests of potential audiences, but it does suggest that, at the least, their issues must be addressed directly in the evaluation and its report.

Evaluation purposes are the reasons for doing the evaluation. What, exactly, will be done with the results? What kinds of decisions are to be made? How will the audiences use the results of the evaluation? Typically, evaluations are conducted to influence decisions about adjustments to ongoing programs, to help determine whether a program should be launched or not (or terminated or continued), to provide a rationale for the continuation of some action (such as the mainstreaming of handicapped children), or to establish accountability for the expenditure of funds or other resources. Prior to launching an evaluation, it is important to determine

its purpose so that the most useful kinds of information can be given to decisionmakers.

Designing the Evaluation

The design of the evaluation will be guided by a series of questions. Each will lead to a series of decisions that move the evaluation activity forward.

1. *What are we evaluating?*

As simple as it sounds, it is absolutely essential to specify, in advance, the precise program or program element that is being studied. This is necessary so that the evaluation is clearly focused on that element and its outcomes, not on ancillary and peripheral issues.

The more specific we can be at this stage, the more precise the evaluation will be. For example, rather than evaluating the activity program, it will be more effective to examine only the intramural component, or the arts component. That way, energy is spent in measured amounts on a topic that will generate interest among a group of people who will then, in turn, be willing to invest energy in it.

Sometimes, more general assessments are necessary. Often, a school wants to determine the effects of major philosophical or mission changes, and, as a result, wishes to look at the school at large. This is quite possible, but it is still necessary to identify, precisely, the components of the school program to be studied: climate, physical space and utilization, instruction, curriculum, guidance. The list is nearly endless.

Schools are very complex institutions, and must be studied in pieces, even if you study all the pieces at once. Labor must be divided, and the most efficient way to accomplish such a division is to give responsibility for each piece to one of the interested parties.

2. *Who are the stakeholders and how can I involve them?*

Who stands to gain or lose something as a result of this

evaluation? More specifically, how can we be certain that the interests of these stakeholder groups are represented on the evaluation?

To begin, it is important to distinguish between two dimensions of evaluation: policy issues and mechanical issues. These two categories can be outlined as follows:

Policy Issues	*Mechanical Issues*
Identification of goals	Organization of evaluation group
Identification of indicators	Facilitation of goal and indicator identification
Selection of measures	Recommendation of measures
Development of analysis plans	Collection of data
Discussion of findings	Analysis of data
Generation of conclusions	Facilitation of discussion on findings and conclusions
Statement of recommendations	Facilitation of discussion on recommendations

Clearly, stakeholders have a very strong, vested interest in what goals will be measured, how they will be measured, how the data will be analyzed, and how the findings, conclusions, and recommendations will be made. It is the job of the evaluator to ensure that these functions are carried out in a credible way.

These role distinctions are not absolute, however. The evaluator may very well be a stakeholder as well, as would be the case if the principal were to lead the evaluation effort. In fact, it is probably most beneficial if the stakeholders are the evaluators; such an arrangement, if the stakeholder group is balanced in its composition, ensures that all possible interpretations of data are offered and that recommendations do not represent a single-value position.

Assemble an evaluation panel. This is a group that will guide the evaluation, make policy issue decisions about the

study, and help interpret the results, conclusions, and recommendations. This panel should be composed of major stakeholders and program representatives. Depending upon the size of the program element being studied, this panel can range from fairly small to quite large, if it is to manage the study of an entire school.

3. *What are the goals of the program?*

Programs often have very specific goals. They may have been specified in the design of the program. Just as often, though, programs have relatively unclear goals because they started out as someone's good idea and simply evolved into their present state.

In the first case, the matter is simple. We simply refer to the original program document for the specification of goals. In the latter case, a bit more work is required.

To identify goals in cases where the original intent was not clear, or when the program has drifted from its original intent, it is necessary to solicit help from those who participate in the program. The easiest way to do this is directly, through interviews or meetings.

In such cases, the evaluator simply interviews a representative sample of the stakeholders in an effort to identify their perception of the program's goals. This approach might be used in our at-risk intervention program cited earlier by asking direct stakeholders (counselors, students, parents, aides) what they believe the program is trying to do and by asking indirect stakeholders (teachers, administrators, other school personnel) what they perceive the goals of the program to be. Note that at this stage no evaluation of the program is being sought; the evaluator is simply trying to achieve consensus on goals.

Following this exercise, a panel of direct and indirect stakeholders should review the goals and attempt to achieve consensus on the most important ones. These goals should be the ones that define the balance of the evaluation.

(Note: At this point it sometimes becomes evident that no consensus exists on the goals for the program. If that is the case, an evaluation is futile. The most productive activity would be to engage in discussions to achieve consensus on the program's purposes. The process outlined above only identifies relatively widely accepted goals when none are specified, when goals are only implied or understood. It does not force consensus where there is none.)

Often, it is appropriate to use some external source of goals, such as are found in the standards of professional associations or in publications such as *Agenda for Excellence at the Middle Level*, or *This We Believe*. * In such cases, these documents outline standards that should be met in a school or its various programs. If there is widespread agreement in the school that these external standards (or some modification of them) are compatible with local purposes, they may serve as a starting point for the evaluation. This, in essence, is the model used by most accrediting bodies and agencies. They provide the standards and it is the responsibility of the school to document that they have met them.

It is essential that all evaluations include attention to the school's mission statement. Schools should not be held accountable for purposes that exist outside their avowed and agreed-upon mission. For a fuller discussion of how to use the school's mission statement for program evaluation, see NASSP's publication, *Developing a Mission Statement for the Middle Level School*, by the Council on Middle Level Education, 1987.

4. *What are the indicators of the goals?*

It is usually impossible to measure goals directly; we must rely on indicators of goal achievement. Because no single indicator is conclusive evidence that a goal has been achieved,

* *Agenda for Excellence at the Middle Level* and *This We Believe* are published by NASSP.

we must usually rely on several such indicators for each goal.

If, for example, we are trying to determine if our intramural athletic program has achieved its goal of improving sportsmanlike conduct, we must examine specific indicators of sportsmanship. We might use a decline in personal fouls as an indicator of achievement of the goal, but we could easily be misled. Perhaps a coach has a very strong position on personal fouls; if you get two, you don't play for a week. This is likely to shape the players' behavior without necessarily changing their attitude toward sportsmanship.

In this case, we need several indicators: absence of personal fouls, congratulation of opposing players for good performance, lack of arguments with officials, and so on. Taken in sum, these indicators will help us determine whether the goal of sportsmanlike conduct has been achieved or not.

Indicators may be quantitative (simple number counts, such as number of fouls, number of congratulations, number of arguments with officials), or more qualitative (opinions of officials about our teams, opinions of opposing coaches, interviews with player to assess sportsmanlike attitudes, etc.). In either case, it is important that they be gathered systematically.

It is also important that the indicators be convincing. In a large suburban school system, one school tried to study the effects of a desegregation program on student attitudes toward different racial groups. They used both paper-and-pencil attitude scales (indicators) and behavioral indicators. Listed below is one of their primary goals and the indicators they used to measure it.

Goal.

- Students will form, voluntarily, racially mixed groups.

Indicators

- Opinions of students: "With whom would you like to work, eat lunch, play on the same team, etc?"

- Behavioral indicators: Racial composition of students in self-selected work groups, racial composition of students in cafeteria and other self-selected environments, racial composition of teams in "choose-up" games or contests, etc.

No single one of these indicators would be convincing evidence that the program was achieving its goal, least of all the opinion survey. However, taken as a group, we can create a convincing argument that they tell us about the status of the specified program goal.

The specification of indicators is crucial and should be a matter for the evaluation panel to discuss. Without agreement on the indicators of goal achievement, there will be no agreement on the outcomes of the evaluation.

5. *What kind of data can we gather about each of our indicators?*

The type of data we gather must be clearly linked to the indicators we have identified. If we agree that student opinions about whom they would select to work with are important, we must be certain that our survey or opinionnaire measures that concept and not something else. If we believe that the racial composition of self-selected groups is a good indicator of positive racial attitudes, we must assess those kinds of groups, not groups that are created by the teaching staff or by other circumstances.

The following section provides a detailed explanation of data types and data-gathering procedures.

6. *How are the data to be analyzed?*

Among the most crucial decisions to be made are those about how the data are to be treated. Three general principles should guide the analysis.

- ☐ The analysis should be in a form that is understandable and useful to the audience.

- ☐ The analysis plan should reflect a broad consensus on the evaluation panel.
- ☐ The analysis plan should be driven by your evaluation questions.

Many lay people distrust the complex data manipulations that are often found in evaluations. Because of that, they tend to find the results of little use in making decisions, and are more likely to rely on preconceived notions and the opinions of others in making judgments about a program and its effects.

The primary responsibility of the evaluator is to ensure that his or her analysis is meaningful to the consumers of the evaluation. In virtually all cases, these consumers are lay people or people with imperfect understandings of complex analytical approaches.

The analysis plan should represent agreement among the evaluation panel members. While you cannot devise an analysis that answers every idiosyncratic question a panel member may have (and you should not try to do so), it is important to determine what kinds of comparisons, contrasts, and descriptions will, in the opinion of panel members, answer the primary evaluation questions.

For example, this is the time to determine if we should compare the students in Program X with those in Program Y; or if we should analyze separately the results for boys and girls; or if we should look at the effects of the program on minority students. These are important issues that are shaped not only by the evaluation questions, but by the political and social context of the community in which the evaluation is being conducted. That is why they must be determined locally.

Finally, the analysis plan should be driven by the original questions posed for the evaluation. Is Program A producing better student achievement than Program B? Are students scoring higher at the end of the program than they did when

they entered? Is the cost of this program in line with programs in other schools, districts, or states? These questions, more than any other considerations, will help the evaluator conduct an analysis that is useful to the consumers.

7. *How can we get consensus on results, conclusions, recommendations?*

Consensus requires deliberation, negotiation, and compromise. That is why, in evaluation, it is important to distinguish between findings, conclusions, and recommendations.

Findings are the actual outcomes of the analysis. Support for findings can be found in the data themselves. One example is that "in special education classes, 23 percent of teacher questions require higher order thinking skills (application, analysis, synthesis, evaluation); 77 percent require lower order skills (knowledge, comprehension.)" This is simply a statement about the data that were gathered. It is not linked to any other data, and it does not carry with it any judgment.

A conclusion is an inference about the data. Conclusions can be relatively simple, such as "special education teachers ask relatively few higher order questions." If findings are combined, however, the conclusions can become more complex (and more highly inferential). If we know that "in regular classrooms 59 percent of teacher questions require higher order thinking skills, 41 percent require lower order skills," we can draw a relatively complex conclusion:

- Regular education teachers ask more higher order questions than special education teachers.
- Our special education students are mainstreamed for half their day.
- Special education students may have difficulty functioning in a regular education classroom because they are asked to think in ways different from what they are accustomed to.

In order to draw this conclusion, we had to know this additional bit of information: students are mainstreamed. If we were unaware of that, we might have missed the potential link between findings *a* and *c*.

Note also that this conclusion is arguable. Perhaps students are not mainstreamed into the classes where higher order thinking is being encouraged. Perhaps the nature of the special education program is to give students the basic skills and information they need to function in a regular classroom, so, of necessity, they focus on lower order thinking skills.

Because of our need to know related information that may not have been included in the study (our students are mainstreamed), and because of the need to consider all possible interpretations of the findings before us, the discussion of conclusions is rightly the job of the analysis panel.

A detailed discussion of analysis strategies follows this section.

8. *What recommendations do we make?*

Again, the recommendations depend largely upon the evaluation questions that were articulated at the outset of the study. Recommendations should answer those questions and provide direction for action.

It is absolutely essential that the recommendations be traceable to the findings and conclusions drawn earlier. If recommendations are not clearly linked to the evaluation data and the group's conclusions, decisionmakers will have little confidence in them.

In the example given above, reasonable conclusions might be to increase the amount of higher order thinking in special education classrooms, to encourage regular classroom teachers to differentiate instruction for special education students, and to provide inservice training for teachers on these concepts. A conclusion that is not justified by these data would be to end the practice of mainstreaming.

Thus, while the positions of all stakeholder groups should

be represented on the evaluation panel and in the recommendation development stage, it is not the job of the representatives to cling tenaciously to a position, regardless of the evaluation data they have helped to gather, analyze, and synthesize. The purpose of representation on an evaluation panel is to participate in the deliberations about the evaluation data, not to defend turf.

9. *How will we disseminate our results?*

Without dissemination, evaluations are useless. They must be widely disseminated to decisionmakers, program managers and personnel, to clients and to other stakeholders. Obviously, the way this dissemination takes place depends upon how the data are likely to be used by each group.

Decisionmakers will need fairly detailed descriptions of processes, outcomes, recommendations, and planned actions. Program managers and personnel will need similarly complete material. Clients (usually students and parents) are usually satisfied with less technical reports, those which outline the general nature of the evaluation and its major conclusions and recommendations.

More remote stakeholders, such as community members with children in school, may be interested in the broadest picture of the evaluation, and what the implications will be for school taxes or other community issues. In this case, accurate news accounts or other public accounts may suffice.

Whatever the forum, it is important for all interested stakeholders to have the opportunity to ask questions about the evaluation and to understand its impact on them. A portion of a board meeting devoted to delivering the report and sketching out its implications for the district may be an important method not only for disseminating results but for gauging public reaction to it as well.

Specific recommendations for disseminating evaluation results are presented later in this monograph.

An Example

The following example shows how, in practice, the model can be used quite simply.

Evaluation Object:

A Program To Improve Student Attributes Toward Mathematics

Goals:

- To improve student attitudes toward the study of mathematics.
- To increase the number of students enrolling in advanced mathematics in high school.
- To increase student involvement in mathematics as a recreational pursuit.

Evaluation Questions:

- Are student attitudes toward mathematics better after participation in the program than before participation?
- Do more students who participated in the program enroll in high school mathematics than enrolled before the program was in place?
- Are students involved in mathematics-related recreational activities to a greater extent than before the program existed? Are more students from the program involved in these activities than students not in the program?

Stakeholders:

Middle level and high school math teachers, parent council, district office

Audience:

Middle level school, high school, math department

Evaluation Panel:

Two middle level and two high school math teachers, one

district office representative, two parents, two students, school administrator.

Goals/Indicators/Data Sources:

Goals	*Indicators*	*Data*	*Analysis*
Improved attitude	Attitude scale	Attitude scores	Compare pre- and post-test scores for both scales
	Subject preference inventory	Subject pref score	
More advanced study	Enrollment in h.s. math courses	Number of enrollees	Compare with enrollment before program
	Enrollment in advanced math by program grads	Number of enrollees	Compare program partic. with non partic
Math as recreation	Membership in math club	Number of members	Compare members before and after program
	Membership in computer club	Number of members	Compare members before and after program
	Math books checked from library	Number of books, number of different titles	Compare before and after program
	Students trying out for math team	Number of students	Compare before and after program
	Placement of school in math olympics	Rank in contest, number of participants	Compare before and after program

From this example, it is clear that the most convincing evaluation results come from multiple data sources. But it is seldom possible, either because of the complexity of the data or the cost of its collection, to use all the data that might be useful. Instead, we must usually focus on a few data sets that will be most useful in fulfilling the purpose for the evaluation. Therefore, we must look for data that meet several criteria:

- ☐ They are clearly related to the indicators and goals.
- ☐ They are relatively easy to collect and analyze.
- ☐ They are likely to be convincing to the consumer of the evaluation.
- ☐ They can be collected accurately and reliably.

Reactive and Nonreactive Data

Two kinds of data meet those needs: reactive and nonreactive or unobtrusive data. Reactive data are those that might actually change the behavior they are trying to measure. For example, if one tries to measure the metabolism rate of a runner, but the instrument used to measure metabolism weighs 18 pounds and must be carried on the runner's back, the extra weight is likely to influence the outcome of the measurement in a significant way.

Among the most common reactive measures are attitude scales. They are reactive because most respondents can figure out what the "right" answer is and give it to the evaluator, regardless of their actual feelings and beliefs. Interviews also tend to be highly reactive, because of their face-to-face nature. While people are not overtly dishonest in answering interview questions, they are not likely to give socially undesirable responses or extreme answers, even if they actually hold those positions.

In short, most tests, surveys, and other conventional data collection devices are reactive. That is not to say that they cannot be useful, though. Reactivity is not usually an issue

unless the test or survey measures some relatively sensitive area, such as feelings about a supervisor, or individual teachers, or behavior that has social judgments tied to it, such as drug use or alcohol use.

Nonreactive measures, on the other hand, tend not to have any impact on the behavior being studied, or at least have a very minimal impact. According to Webb's book, *Unobtrusive Measures,* there are five categories of nonreactive or unobtrusive measures.

Physical Traces

Most behavior leaves some form of physical trace, either by wearing something down (erosion) or building up some kind of residue (accretion). It is easy to detect which magazine in the middle school library is most popular, because it is in tatters by the end of the month. Most of us can also tell which lunch line (a la carte or full lunch) is most popular by the number of people standing in it. There are literally hundreds of these data sources.

The most famous is probably a study done by a museum director to measure the intensity of interest shown in a display by very young children. At predetermined times, he would go to the display, count the finger and nose prints on it, and then wipe the glass clean, only to return again to count the prints. After a period of time, he was able to gauge the amount of interest shown in the displays by the accumulation (accretion) of smudges on the glass.

Archives I. The Running Record

All institutions keep records as a matter of course. Many are required by law, such as attendance, grades, and permanent files. Others are kept as a matter of district or school policy, such as discipline referrals, requests for transfers by teachers, and athletic physical records. These records can yield very useful information about an institution, information that is often very helpful in conducting an evaluation.

One principle that should guide all data collection is that if the data already exist somewhere, don't collect them again. An astute evaluator will consider the permanent records of the institution a good source of information.

Archives II. The Private, Episodic Record

In addition to official records, most individuals keep records of their own, sometimes no more formally than saved phone messages or notes scribbled on a steno pad. Many principals keep records on disciplinary referrals and the actions taken; teachers will often keep notes about parent or student conferences. These private records can provide useful insights about a program, sometimes in surprising ways.

One principal noted that, following any evening activity at the school, the number of calls from parents expressing concerns about the way the event was managed increased dramatically if there were not many parents in attendance at the event. He noted this by looking at the telephone log for the week following the event, and comparing the number of complaints received after events many parents attended with the number in the week following a less-well-attended event.

Observation I (Direct)

Direct observation is conducted openly. In this case, the observed may or may not know what, exactly, the observer is looking for, but they are generally aware that they are being observed, or at least watched.

How can this be unobtrusive? In some cases, observers are natural parts of the landscape, such as at an athletic contest or artistic performance. In such cases, one observer, more or less, makes little difference in the performance of the task or act.

At other times, such observation may be included as part of the normal routine of an event. This would be the case where supervision is the norm (such as in a lunchroom or a library); the students are accustomed to being watched while they are in those places.

It is an easy matter to make the "watching" more systematic: to see how many mixed-sex tables there are at each grade level lunch, or to determine how much sharing of material goes on during a library research project.

Observation can also be unobtrusive if the observation cannot influence the event. Watching a youngster run the 100-meter dash is not likely to affect her time in the run. The performance of some learned skill or some kind of artistic or athletic event is not likely to be altered, substantially, by the presence of an observer.

Observation II (Contrived)

Contrived observations occur when the observer remains anonymous, often when the observation is conducted through some electronic medium, such as recordings. Sometimes these observations are conducted through the famous one-way mirrors that many of us remember from our days in teacher training. More often, the observation is conducted on recorded segments, either video or audiotapes.

Another form of contrived observation is when the subject has no clue about what the observer is looking for, or if the subject is told that the observer is watching one thing when he is, in fact, watching something else.

In one study, parents and young children were put in a room with a number of toys at one end, although not in sufficient numbers for each child to have one. The parents were told that the lab assistants in the room were observing the kind of sharing behavior that took place among the children. In fact, they were observing the interactions between children and their parents, as the parents encouraged (or did not encourage) their children to share the toys.

Such clandestine observations are fraught with ethical problems and are best avoided. Recorded observations, on the other hand, made with the knowledge of the subject(s), are often useful as a way of "visiting" many classrooms in a short period of time. These recordings have the added advan-

tage of preserving the experience so it can be analyzed more carefully.

Obviously, these measures will not withstand severe scrutiny. But, when they are used in combination with other, more familiar measures, we can achieve our goal of employing multiple measures without the need to invent yet another survey to correlate with the one we may have just administered.

The example given earlier indicates a number of cases in which both reactive and nonreactive measures can be used to give a more complete picture of a mathematics program. In fact, only two of our proposed data sources (the survey and the subject preference inventory) are "reactive" in the strictest sense of the word. All the other measures make use of observations and both official and unofficial records.

Prepared Evaluation Systems

A number of evaluation systems have been prepared for use in middle level schools, but this number remains fairly small considering the number of middle level schools in the United States. The primary reason for this shortage seems to be twofold:

- There is no consistent, concrete model for middle level education in the United States; and
- Middle level schools seem to grow as local phenomena, due primarily to the lack of clear state or national standards for accreditation.

In many respects, these may be seen as strengths of the middle level school as it exists in this nation. The schools are intimately connected with the communities they serve, and

they respond to the needs of the local environment and the children in the community rather than an abstract listing of "shoulds" or "musts." Nevertheless, such fluidity makes the evaluator's job somewhat more difficult.

While this document does not purport to be a catalog of such systems, several are discussed here as examples of the genre.

One evaluation system, the Middle Grades Assessment Program (MGAP), is based on research on effective middle level schools (Lipsitz, 1984), and, on the basis of that research, outlines program and organizational conditions that enhance learning for young adolescents. It is a very comprehensive view of a school, and it produces huge amounts of data on virtually every facet of school operations.

Most important, it allows (in fact, it requires) evaluators to examine the interconnections that exist among various phases of the program, and it assesses the ways in which each program element supports nine desirable conditions for middle level education: safety, academic effectiveness, diversity, self-definition, participation, social interaction, physical activity, competence and achievement, and structure and limits.

A somewhat shorter assessment program is published by the New England League of Middle Schools and is based on the statement of philosophy adopted by the National Middle School Association, *This We Believe*. In this assessment, indicators are provided under each of the 10 essential elements of middle level education outlined in the source document.

Evaluation guides are also available from most of the regional accrediting agencies, such as the North Central Association of Schools and Colleges. These agencies use a standard publication, *Evaluation Criteria for Middle and Junior High Schools*, published by the National Society for the Study of Evaluation. It is the most widely used instrument for middle grades assessment, particularly as the evaluations are conducted for the purpose of accreditation.

This publication gives very broad categories of criteria,

most of which are connected with the criteria for accreditation articulated by the various regional accreditation associations. As such, this document tends to focus on program standards and the most desirable conditions for middle level schools.

All these systems are useful, provided they are modified for local use. It is important that evaluations not only respond to some larger vision of what the school might be, but to the local culture that nurtures the school as well. Indicators can and should be added to each of the major categories outlined in these prepared systems, for schools have unique ways of addressing the needs of their students. In any event, these standards systems are good models to follow provided they are not treated as sacred text and not to be altered.

Data Analysis and Interpretation

Truth, if not beauty, is in the eye of the beholder. Or, as one sage put it, "where you stand depends largely on where you sit."

These two bits of wisdom are offered as an introduction to this section because the analysis and interpretation of data is one of the most difficult and emotionally charged aspects of an evaluation. "What does it all mean?" is the question to be answered. And everyone who has a stake in the answer to that question is likely to have an opinion about what the answer should be.

Often we tend to skirt this particular issue by trying to design evaluation systems that are "foolproof" in terms of interpretation. Ideally, under such a system, we plug in the numbers (or other data) and the answers roll out of the machine. But like most ideals, this one doesn't exist either, and the evaluation studies that produce clear-cut results are

probably studying something so trivial that no one is likely to have a strong opinion about it, anyway.

Two other tendencies may exist as well. We either "reify" human events with numbers, or we dazzle people with complex analyses. Reification is the process of treating an abstraction as if it is real in some concrete sense.

For children, we use folk figures, fables, and other tales to reify abstract values that we want them to learn, and the practice is useful. For researchers, it can be a trap, for once we have assigned a value to an abstraction, we believe that we can treat it like all other abstractions to which we have assigned that same value.

For example, the normal midpoint response on a Likert Scale is "undecided" or some other neutral expression. Suppose a group of teachers, when given the stimulus statement, "My principal exerts instructional leadership," reply "undecided." It would be tempting to conclude that all of them have given the question some thought and cannot find enough evidence one way or the other to answer the question.

That, however, is only one plausible reason for selecting that answer. One teacher may select it because he has no idea what "instructional leadership" means as used in this sentence. Another may be undecided because he really doesn't see much evidence of that behavior in the principal, but, because he likes her, he gives her the benefit of the doubt and uses a neutral response. Still another may be indicating that she is not certain principals should engage in instructional leadership, and yet another may mark "undecided" to this question (and all the rest as well) because he couldn't care less about the survey.

Those are at least five plausible reasons for selecting that response, none of them enough like the others to say that they all belong together, yet that is precisely how we would treat them in this kind of measure.

This is not to say that these scales are without value. It simply reinforces the need to gather data from multiple sources, using a variety of measures.

Fancy analytical footwork in an evaluation often conceals a poorly framed question or data of questionable quality. The analysis of data should be as simple as possible to accurately address the question being answered. It should not be complicated by unnecessary, peripheral questions, or attempts to control all the extraneous variables that exist in a school.

Analyses should be as complete as possible, recognizing the limits of the school environment for data collection and the limits of lay (and many professional) audiences in interpreting complex statistics.

With all these limitations, how can an analysis be conducted? The answer is quite simple: the analysis, like the evaluation itself, is essentially a social process. It requires negotiation, deliberation, argument, evidence, and ultimately, compromise. No one has the corner on objective truth in an evaluation; it is the group that must make meaning out of the data amassed before them. This process can be eased by several steps, as outlined below.

1. Reduce the data.

No one can deal effectively with all the information presented in raw form. The first step must be to put the data together in some selective way so that it is meaningful to the interpreter. For most numerical data, this means simply conducting the analysis called for in the original plan (comparing the enrollment before the program with the enrollment after the program).

Most often, this is accomplished with basic data descriptions, such as measures of central tendency (means, medians, modes) and variability (variance, standard deviation). Simpler presentations may be made with frequency graphs, histograms, scatter plot charts, or a host of other graphic display methods.

Generally, this step is followed by the actual statistical analysis of the data, if one is planned. This would include comparisons of data sets (such as pre and post-tests or "before" and "after" measures) with simple inferential statistics: t-test,

analysis of variance. It may also call for the analysis of frequency data (with chi-square analyses) or relationships among data sets (with simple correlations). In all cases, the methods to be used should be derived from the questions to be answered, not from a pre-ordained statistical plan that will answer standard questions.

2. State findings and observations that come from the data.

Findings and observations are statements that can be made directly from the data before us. No inferences are necessary, no explanations are offered, and no judgments are made.

For example, if the scores on the attitude scale administered after the math enrichment program are significantly higher than those administered before the program, we would state, simply, "Attitude scores collected after the enrichment program were higher than those collected before the program was run."

If a comparison of boys' and girls' grades in physical education were conducted and the chi-square analyses were to be significant, we would state, "The distribution of physical education grades for males and females is significantly different."

The purpose of the statement is to explain, very clearly, what the data mean. Once the analysis group has listed these statements, it is ready to begin drawing conclusions.

3. Draw conclusions—Phase I.

In Phase I of this process, the analysis group looks at the list of findings and tries to identify patterns that allow us to draw meaningful conclusions from the information. Conclusions are inferences that can be drawn from the information before us.

If, for example, the analysis team identified the following findings, they might draw conclusions such as those shown in the box.

Findings	Conclusions
1. Attitude scores collected before the math enrichment program were lower than scores collected after the program.	1. The enrichment program improved student attitudes toward mathematics.
2. On a subject preference inventory administered after the enrichment program, math gained one rank over the same scale administered before the program.	2. The enrichment program improved math's "standing" in the preference of students in the enrichment program.
3. Enrollment in the Math Exploration Club increased by 22 percent from the beginning of the enrichment program to the end of the program.	3. The enrichment program increased student interest in mathematics as a recreational pursuit.
4. Eighth graders participating in the enrichment project were 50 percent more likely to enroll in advanced ninth grade math than eighth graders not participating in the project.	4. Eighth graders who participated in the project are more likely to select mathematics study at the high school level.
5. Girls' mathematics grades did not increase significantly between the beginning and end of the project; boys' grades did.	5. The enrichment project has a more favorable impact on boys than on girls.

It's easy, isn't it? Most of the time it is, as indicated by numbers 1, 2, and 3 in the box. Those findings are pretty clear, and the conclusions quite obvious. But the findings for items 4 and 5 are not so clear and not so obvious. There are several alternative interpretations for the findings, and the conclusions listed may not be the best possible ones.

In conclusion 4, it may be that only the most able students

participated in the math enrichment program; thus, it would make sense that this group would be more likely to take advanced math. In other words, the finding may have come from who was in the program rather than what the program did.

There are a host of possible interpretations for conclusion 5. It may be that girls, as a group, had higher grades to start with and are thus not likely to make significant progress. It may also be that traditional biases toward women in mathematics crept into the enrichment program and turned some of the girls off to its benefits. In either event, the conclusion that was drawn may not be entirely justified.

In this phase of the conclusion-drawing process we are interested only in the conclusions that come, clearly and naturally, from the data. In our first pass through the findings, we list only the conclusions that can be drawn with reasonable certainty and that enjoy widespread agreement among the people on the analysis team.

4. Draw conclusions—Phase II.

In this phase, we deal with conclusions that are not as obvious as the ones that became clear in our first review of the findings. This process requires discussion, negotiation, and often, compromise.

For those findings where the conclusion is not clear, it is important to start with a very open discussion of all the plausible reasons for the finding to be evident. In some respects, this is a brainstorming session, although it is constrained somewhat by the data before us.

After that, the group can probably eliminate many of these potential explanations because they are simply outrageous or not supported by the data. Ironically, this process can actually be fun.

After the list has been shortened as described above, create subgroups within the larger analysis group and assign each the task of drawing conclusions from the findings that: re-

spect the data on which the finding was based; and incorporate as many of the group's explanations as possible.

Following this process, the entire group reviews each subgroup's conclusions and selects the ones that are most representative of the entire group's opinions, modifying the conclusion as necessary.

This process takes care of most disagreements, provided the group remains focused on the data that they have collected and does not go beyond those data.

It is better to draw modest conclusions rather than extravagant ones, since modest conclusions tend to reflect the data more accurately. As an example, instead of saying that boys benefited from the enrichment program more than girls (conclusion 5), it might be better to say, "The program had a differential effect on the grades of boys and girls enrolled in it."

In cases where it is not possible to draw firm conclusions, offer multiple, plausible conclusions with the admonition that the data were insufficient to answer the question.

Taking as an example finding 4, we can see that there are several plausible explanations. First, the program did have an effect on high school enrollments of the students who participated. Second, the program attracted a certain type of student who was more likely to enroll in advanced math in the first place. Or, third, that some other effect we cannot even imagine was operating.

In this case, the group might elect to gather more information, this time on the composition of the groups participating and not participating in the enrichment project. Such data will help to reinforce or eliminate one of the competing explanations, and make it more likely that the group can draw a conclusion.

In cases where it is not possible to do additional investigation, either because of lack of time or lack of resources, it may be necessary to simply state the competing explanations for a finding and move along from that point.

Use of Questions and Analytical Frameworks

Once the list of findings and conclusions has been drawn up, it is time to turn our attention back to the original questions that framed the evaluation. If the original questions were clearly stated, this becomes a relatively easy process.

First, we list the questions we set out to answer. These are usually taken from the goals and objectives statements on our evaluation system plan (see page 17). For example, in the case of our mathematics example, the questions include:

- Does the program improve student attitudes toward math?
- Does the program increase the likelihood of students enrolling in advanced high school mathematics?
- Does the program increase student participation in mathematics-related recreational activity?

Now it's time to put it all together. Using the questions as an outline, we must prepare answers to them based on the list of findings and conclusions we have gathered and determined. For the sake of objectivity, it is important that members of the analysis team participate in every phase of this activity.

Once again, we use subgroups. Each subgroup is assigned a number of questions to answer. At the same time, we ensure that every question is being addressed by at least two different subgroups, working independently of one another, but from the same list of conclusions and findings.

Thus, our writing phase may look like this for four groups:

Question	Group	Group
Does the program improve student attitudes toward math?	1	2
Does the program increase the likelihood of student enrollment in advanced math courses?	3	4
Does the program increase the amount of student participation in math-related recreational pursuits?	1	2
Does the program increase student interest in math-related careers?	3	4

After each subgroup has written a response to their questions, they meet with the other subgroup assigned the same task and develop a joint statement. This process is repeated for each question, thus ensuring ownership by the entire group in the report that is forthcoming from the evaluation team.

Dealing with Unanticipated Findings

Often something emerges from the data that was not part of the original investigation, but which is so important, either to the program or the school, that it must be included in the evaluation report.

One such instance occurred in a school located in a state that had implemented a "no pass, no play" policy for extracurricular activities. Students were required to maintain a B average to remain eligible for participation in special activities. A study of the attrition rate between middle level and high school science enrollments revealed that a large number of good students were not pursuing Advanced Placement and

more advanced sciences because the difficult grading standards jeopardized their chances of participating in extracurricular activities. Rather than run the risk of being excluded from this participation, the students simply avoided more difficult courses.

In this case, the conclusion to be drawn from the finding that students were avoiding advanced study in science had little to do with the objectives, or the science program, or the way the program was operated. Instead, it raised a schoolwide issue that resulted from a state policy mandate. This unanticipated finding helped illustrate the impact of these mandates on day-to-day operations.

Unanticipated findings should be reported both in the final report of the evaluation team and to the specific segment of the school that is most directly affected. In the example cited above, the science department not only included the finding in their report, but shared it with the superintendent and school board, and, through them, with the state superintendent of public instruction and area legislators. The information became significant in leading the legislators to an understanding of the consequences of one of their policy decisions.

Making Recommendations

Ultimately, evaluations must lead to recommendations for specific action if they are to be useful. At this point, conditions other than the evaluation itself enter into the decisions. The availability of resources, a political climate that will support change, and the priority assigned to the program by the institution as a whole all operate to constrain both the number and nature of recommendations the evaluation panel is likely to make.

Several dimensions must be considered in the formulation of recommendations: priority, resource demand, and feasibility.

Priority

The priority of any given issue is dictated by its relation to the central goals and mission of the institution in which it arises. Issues that are very close to the central purposes of the institution are addressed first; those that are further from the primary mission of the school get somewhat less attention.

It is critical that the evaluation panel recognize the extent to which the issues they have uncovered and the conclusions they have drawn relate (or can be made to relate) to the central objectives of the school or school district.

No one is likely to care very much if students are not wild about mathematics. However, if it is shown that student attitude influences their decision to pursue advanced study in math, thereby affecting their subsequent college or occupational opportunities, the issue becomes central.

This is not to say that evaluation panels should draw spurious links between their findings and the goals of the school simply to get attention. It does suggest that the evaluators must be astute about the principal mission of the school and know how to cast the evaluation results in terms of that mission to the extent it is possible to do so.

Recommendations, then, should be inexorably linked with the overall purpose and mission of the school in order to raise their priority status in the institution and increase the likelihood that they will generate meaningful action.

Resource Demand

The concept of resource demand deals with how much a recommendation will cost in terms of money, personnel, time, and effort. Recommendations that carry relatively low resource demands are more likely to be implemented than those with higher costs. In fact, institutions are often tempted to act

on several lower-priority recommendations because they cost relatively little, rather than on a higher priority, more expensive demand.

Evaluators must be sensitive to probable costs of their recommendations, particularly if such recommendations result in large-scale capital expenditures. In such cases, the evaluators must be prepared with data quite different from the kind normally collected in school studies, or they must be prepared to make somewhat more modest recommendations.

The study of a physical education program may result in several recommendations for the purchase of new equipment, a change in scheduling to better utilize the specialized facilities, and reconstructing the ventilation system in the gymnasium.

In the first two instances, boards of education are inclined to accept the judgment of educators, since that is their field of expertise. No matter how much their judgment is trusted, though, no board is likely to engage in a major capital expenditure on the basis of a program evaluation study. Much more specialized investigations are needed.

In such cases, it would be wise for the evaluation panel, rather than recommending the reconstruction of the ventilation system, to recommend that a study be undertaken of the air quality in the gymnasium and the efficiency of the ventilating system. That recommendation, calling for the study by experts, is much more likely to be implemented, with the desired result that ventilation be improved.

With a less modest recommendation, the focus of the debate may shift to the qualifications of the panel to make such recommendations, and the need for new equipment and new schedules, matters on which the panel is indeed expert, will be lost in the larger discussion.

It is also important to note that recommendations for expenditures that support existing investments are received with more interest than those which require discarding an existing investment. For example, district administrators may be more

interested in a recommendation for upgrading the existing equipment in the computer laboratory rather than abandoning the present hardware in favor of new equipment, no matter how much more beneficial that new equipment is likely to be. Consideration should be given to the way in which such recommendations are framed.

In short, the recommendations that are costly must be made from impeccable data sources, and must be in areas where the evaluation panel's expertise is unchallenged. Further, such recommendations are likely to be funded only if they support the central mission and goals of the school, and the links are drawn clearly and accurately for the decisionmakers.

Feasibility

The feasibility of any given recommendation has an important bearing on its adoption and implementation. While feasibility does deal with both priority and resource demand, it has other dimensions as well.

First, feasibility is a function of displacement or modification. What will be displaced or changed by this recommendation? If a proposed exploratory music program is going to upset the offering of a very effective and very popular instrumental music program in the school, the proposal has little chance of success.

By the same token, a proposed after-school activity program that is suggested because none exists has a good chance of success. Thus, it is important to identify the kind of displacements that might occur as the result of a recommendation and be prepared to address those displacements directly.

Second, feasibility is related to the political climate of the school and the community. If a parent group has fought long and hard for a pull-out gifted and talented program, a recommendation that the program be abandoned is not going to be very widely acclaimed. In fact, such a recommendation invites active, organized resistance.

At the same time, a recommendation for improving the quality of the life fitness component of the physical education program, which coincides with the release of a major study on the declining fitness of American youth, may meet with great success. Evaluators must be sensitive to the political and social milieu in which their recommendations are offered.

Third, feasibility is clearly related to the reward and benefit being offered by any recommendation. One of the first questions asked by most people when presented with a recommendation is, "What's in it for me?"

That's not selfish; it's simply reasonable to assess how a proposed change will affect you, your child, your job, or your working conditions. Recommendations that promise only a loss cannot be expected to be implemented; or, if implemented, will not enjoy wide support.

For example, in one school a recommendation was made to do away with department heads in the middle school, replacing them with team leaders. Some of the teachers in specialized areas argued that a team leader did not replace a department head, and it took someone knowledgeable about their programs to keep equipment working, supplies stocked, textbooks and other instructional material up to date, and relationships open with the high school subject area teachers.

Dissatisfaction remained for a long time after the recommendation was implemented, not only because it was viewed as a loss to several groups of people in the school, but because the "replacement" was not valued as highly as what was given up.

Evaluators must be very sensitive to the impact their recommendations are likely to have, and be sure that any recommendation does not contain the seeds of its own failure. Most often such failure results from ignoring the kinds of things that will be displaced by the new recommendation, not attending to the political climate of the school and community, and failing to consider "What's in it" for the various stakeholders.

Dissemination of Results

Before launching an evaluation, the panel should determine a dissemination plan. Dissemination plans contain two major items: targets and methods.

Dissemination targets are those who have a need to know the outcomes of the evaluation. For the most part, this includes all the stakeholders in the community, both direct and remote. For each group, the evaluators should be able to state that group's "need to know."

Need to know is an estimate of the importance of the information to any individual or group. Generally, those with the strongest need to know are those:

- Who must make a decision about recommendations from the evaluation
- Who must alter their work behavior as a result of the evaluation
- Whose children will be directly affected by the results of the evaluation.

Clearly, this includes most of the stakeholders included on the original evaluation panel described at the beginning of this monograph.

Beyond this group is a group with a right to know. This group includes:

- Taxpayers or others expected to fund the recommendations
- Others who work in the setting where changes will be made, even if their job is not directly affected
- Those whose confidence in or contact with the school may be altered as a result of a program change (potential

employers, youth agencies, libraries, and a host of other individuals and agencies who come in contact with the children from the school).

While these individuals have a right to know, that right will be satisfied with less detailed information that addresses the recommendations and outcomes of the evaluation rather than its methods and procedures.

Once needs and right to know have been established, it is a fairly simple matter to develop dissemination plans that reach each audience. Obviously, a full, detailed report should be given to decisionmakers and those whose work is affected by the evaluation. A somewhat less detailed report, listing major findings, conclusions, and recommendations may suffice for others in the "need to know" category.

For individuals in the "right to know" group, much shorter and less detailed reports are appropriate, such as those commonly found in newspaper reporting. In fact, the media offer a good way to disseminate these findings, particularly if the evaluation panel provides a clear, straightforward press release that lists the major points of the evaluation. In such cases, copies of the full report should also be made available to the community members, either by loaning them from the schools or by placing them in the public libraries or community centers.

Ultimately, just as important as the quality of the study are the pains that are taken to be sure the results are widely known. As in all cases, though, good judgment should prevail. While the community at large (and, therefore, the press) may be interested in the assessment of a school district's literacy program, they are probably somewhat less interested in the status of materials for teaching handwriting at the middle school level.

In fact, it would be foolish to overload the communication system with trivial details, since it will make the press and other information carriers less sympathetic to carrying stories

about larger issues when the need arises. Generally, the community will be satisfied with accurate summary information that signals the availability of more in-depth information if such detail is needed.

Some Pitfalls

Because of the complexity of evaluations, a significant number of things can go wrong. However, most of the things that do go wrong are technical in nature and can be easily corrected. More difficult problems occur when something goes wrong with the evaluation system. Normally, these problems take one of several forms.

Personnel changes often cause evaluations to stall or to grind to a complete stop. This occurs most frequently when the individual directing the evaluation transfers or resigns. Generally, this occurs only in circumstances where there has not been a concerted effort made to involve a broad base of stakeholders in the process. This problem, then, can be prevented more easily than it can be solved. The prevention is simple: Involve enough people at the outset so that the process is not driven by one person.

Resistance to evaluations usually arises as a result of fear. Some individuals fear the effects that the evaluation will have on their programs or even their jobs. The resistance can range from a failure to cooperate to outright attempts to sabotage the evaluation process.

Once again, prevention is the best approach. Involving all interested parties in the early stages of planning will help to overcome the fear that the evaluation is driven by external forces over which individuals have no control.

Becoming overwhelmed with information is a real problem.

The tendency is to collect more information than evaluators believe they need, "just to be safe." This is usually accompanied by the belief that we can just ignore data that we don't need. Unfortunately, whenever data are collected, a contract of sorts is established between the evaluators and the subjects.

It goes something like this: "If I take the time to give you this data (or respond to your requests for information), I expect you to use it." If the evaluator breaks that trust, he or she can expect negative reactions to the accuracy of the study ("It seems like you laundered the data") and little cooperation in the future. The lesson here is simple: Gather enough data, but only enough.

Loss of commitment to the evaluation comes when people get into the process, then see just how time-consuming and demanding it can be. At this point, there is frequently a loss of commitment on the part of the evaluation team. This can be resolved in several ways.

First, involve people who have a stake in the outcome of the evaluation. Their commitment is likely to remain high. Second, "chunk" your evaluation into manageable units, one program or activity at a time, rather than the whole district or the whole school. Finally, be sure that the evaluation is perceived as realistic and meaningful. Get commitment that the results of the evaluation will be used. No one wants to work on a project that will just be shelved.

Ignoring the results of the evaluation takes two forms. Either the governing body (a board, for example) gets enough preliminary information to believe they know exactly what is going on, or they simply decide that the evaluation does not lead in the direction that they want to go. This is the most difficult problem to deal with, but some measures are possible.

First, the evaluation team should get a commitment from the governing body (whether it is a board of education, a steering committee, or the principal) that a response, preferably in writing, will be made to the results of the evaluation.

CO

In short, this is a guarantee that the governing body will pay attention to the findings.

Second, the evaluation team should build a wide base of support for the evaluation among the stakeholders and in the community. If many people are aware of the evaluation, it is unlikely that it can be ignored.

A Final Note

Evaluation is no mystery. It is simply a process whereby judgments are made on the basis of information that is gathered in systematic, convincing ways. Without systematic assessment, schools become the victim of fads, whims, and vocal pressure groups. With it, school improvement can proceed judiciously and effectively.

References

Dorman, Gayle. *Middle Grades Assessment Program.* Carrboro, N.C.: Center for Early Adolescence, 1984.

Lipsitz, Joan. *Successful Schools for Young Adolescents.* New Brunswick, N.J.: Transaction Books, 1984.

nassp
1904 Association Drive
Reston, Virginia 22091-1598